On Utilizing Association and Interaction Concepts for Enhancing Microaggregation in Secure Statistical Databases

B. John Oommen, *Fellow, IEEE*, and Ebaa Fayyoumi

***Abstract*—This paper presents a possibly pioneering endeavor to tackle the Microaggregation Techniques (MATs) in secure statistical databases by resorting to the principles of associative neural networks (NNs). The prior art has improved the available solutions to the MAT by incorporating proximity information, and this approach is done by recursively reducing the size of the data set by excluding points that are farthest from the centroid and points that are closest to *these* farthest points. Thus, although the method is extremely effective, arguably, it uses only the proximity information while ignoring the mutual interaction between the records. In this paper, we argue that interrecord relationships can be quantified in terms of the following *two* entities: 1) their "association" and 2) their "interaction." This case means that records that are not necessarily close to each other may still be "grouped," because their mutual interaction, which is quantified by invoking transitive-closure-like operations on the latter entity, could be significant, as suggested by the theoretically sound principles of NNs. By repeatedly invoking the interrecord associations and interactions, the records are grouped into sizes of cardinality "k," where k is the security parameter in the algorithm. Our experimental results, which are done on artificial data and benchmark real-life data sets, demonstrate that the newly proposed method is superior to the state of the art not only based on the Information Loss (IL) perspective but also when it concerns a criterion that involves a combination of the IL and the Disclosure Risk (DR).**

***Index Terms*—Information loss (IL), interaction between microunits, interrecord association, microaggregation technique (MAT), secure statistical databases.**

I. Introduction

MUCH attention has recently been dedicated to the problem of maintaining the confidentiality of statistical databases through the application of statistical tools to limit the identification of information on individuals and enterprises. Statistical Disclosure Control (SDC) seeks to balance the confidentiality and the data utility criteria. For example, federal agencies and their contractors who release statistical tables or microdata files are often required by the law or by established policies to protect the confidentiality of released information. However, this restriction should not affect public policy decisions, which are made by accessing only nonconfidential summary statistics [2], [43]. SDC can be applied to information in several formats, e.g., tables, responses to dynamic database queries, and microdata [8], [13], [15], [19], [22], [35], [37], [38], [42], [52], [63]. The protection that SDC provided results from either generating a set of synthetic data from a model that was fitted to the real data or modifying the original data in a special way [8], [14], [30], [63], [64].

Microaggregation is one of the most recent techniques that has been used to mask microindividuals in terms of protecting them against reidentification in secure statistical databases [6], [13], [24], [40], [49], [55]. Moreover, it is modeled as a clustering mechanism with group-size constraints, where the primitive goal is to group a set of records into clusters with, at least, size k based on a proximity measure that involves the variables of interest [10], [20], [29], [39], [44], [55], [60].

The Microaggregation Problem (MAP), as formulated in [10], [24], [39], [44], [49], can be stated as follows. A microdata set $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ is specified in terms of the n "microrecords," i.e., $U_i's$, with each representing a data vector whose components are d continuous variables. Each data vector can be viewed as $U_i = [u_{i1}, u_{i2}, \ldots, u_{id}]^T$, where u_{ij} specifies the value of the jth variable in the ith data vector. Microaggregation involves partitioning the n data vectors into m mutually exclusive and exhaustive groups to obtain a k-partition $\mathbb{P}_k = \{G_i | 1 \leq i \leq m\}$ such that each group G_i of size n_i contains either k data vectors (a fixed-size case) or between k and $2k-1$ data vectors (a data-oriented case).

The optimal k-partition $\mathbb{P}_k^*$ is defined to be the one that minimizes the within-group dissimilarity that is given by the *sum-of-squares error*, i.e., $\text{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^T (X_{ij} - \bar{X}_i)$. This quantity is computed based on the Euclidean distance of each data vector X_{ij} to the centroid $\bar{X}_i$ of the group to which it belongs. The *Information Loss* (IL) is measured as $\text{IL} = \text{SSE}/\text{SST}$ (and is sometimes specified as a percentage), where SST is the squared error that would result if all records were included in a single group and is given as $\text{SST} = \sum_{i=1}^{n} (X_i - \bar{X})^T (X_i - \bar{X})$, where $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$.

As mentioned in the literature, this problem, in its multivariate setting, is known to be *NP*-hard [53] and has been tackled using different approaches, e.g., hierarchical clustering

Manuscript received July 18, 2008; revised March 11, 2009. First published July 28, 2009; current version published October 30, 2009. This paper was presented in part at the *Ninth International Conference on Information and Communications Security*(*ICICS* 2007), Zhengzhou, China, December 2007 [54]. This paper was recommended by Associate Editor P. S. Sastry.

B. J. Oommen is with the School of Computer Science, Carleton University, Ottawa, ON K1S 5B6, Canada, and also with the University of Agder in Grimstad, 4876 Grimstad, Norway (e-mail: oommen@scs.carleton.ca).

E. Fayyoumi was with the School of Computer Science, Carleton University, Ottawa, ON K1S 5B6, Canada. She is now with the Department of Computer Information System, Faculty of Prince Hussein Bin Abdalla II for Information Technology, Hashemite University, Zarqa 13115, Jordan (e-mail: efayyoum@scs.carleton.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2009.2024949

[24], [48], [49], genetic algorithms [24], [48], [49], [59], graph theory [39], [44], fuzzy clustering [31], [61], and machine learning [34]. All the heuristic Microaggregation Techniques (MATs) seek to minimize the value of the IL. However, minimizing the loss in the data utility is an important issue, which is difficult to enforce, primarily because this strategy was intended to enhance the security in an SDC technique. Apart from the IL, in this paper, we contend that the loss of *confidentiality* as a result of disseminating the published microaggregated file must also be analyzed. This case is true, because the so-called Disclosure Risk (DR) depends not only on the data but also on the intruder who knows something about the population.[1] Indeed, the definition of optimality for an SDC is defined in the literature as being equivalent to offering the best tradeoff between the IL and DR [12], [50], as advocated in recent studies [27], [32]. Therefore, any DR assessment must evaluate the risk of providing additional information that can assist in linking a masked record with the corresponding record in the original data set. Although the formalization of the DR will be done later in this paper, at this juncture, we mention that we intend to present a scheme that addresses both the IL and the DR criteria.

As argued by most researchers, maintaining a happy medium between the IL and DR is not trivial. Indeed, even the question of how such a compromise can be attained has not been fully investigated. In this paper, we shall argue that a good MAT can minimize the IL and yet attain a suitable value for a well-defined composite measure. One such measure, which we refer to as the Scoring Index (SI), is a linear combination of the functions of the IL and DR. Our aim is to find a good strategy for optimizing the SI.

In general, minimizing the IL directly follows maximizing the similarity between records in each group. The state-of-art MATs depend on utilizing the "Euclidean" distance, which serves as the criterion that plays a central role in estimating the similarity between the records. However, this distance function does not completely capture the appropriate notion of similarity for any data set. Our position is that the notion of similarity should be measured using a metric that also unravels the relationship between the interrecords. We believe that this approach can be quantified in terms of two quantities: 1) the mutual "association" between the individual records and 2) their mutual "interaction." We propose to measure these quantities using Association Similarity Rules (ASRs), which are well-known data mining techniques for discovering the relationships between patterns in different application domains [3]–[5], [36], [46], [56]. In this context, we mention that the concepts of association and interaction are derived from the Associative Cluster Neural Network (ACNN), which estimates the similarity between neurons by building a dynamic model that was evaluated through the interaction between the neurons inside each group and the interaction among the groups themselves. The main contribution of this paper is to integrate the basic concepts of ASRs with MATs to devise a new strategy for estimating the similarity. This new method demonstrates that the IL can be reduced by taking two measurements into consideration. First, we consider the mutual association between the records. Second and more importantly, we consider the mutual interaction between the records by using a transitive-closure-like operation when $k \geq 3$. This approach, in turn, is achieved by invoking our newly proposed Interactive–Associative Micro-Aggregation Technique (IAMAT). Indeed, the proposed scheme, i.e., IAMAT, is a variation of Maximum Distance to Average Vector (MDAV), which uses association and interaction, instead of the Euclidean distance, as measures to aggregate the records into groups. The effect of these considerations is shown to minimize the IL by up to 13% compared to the state of the art. In addition, IAMAT yields the best reported values for the aforementioned index, i.e., the SI. We argue that the applicability of the new strategy in estimating the similarity provides a promising strategy to effectively protect sensitive data in the microdata file based not only on minimizing the value of the IL but also on offering the best tradeoff between the IL and the DR.

[1]Note that the DR in an MAT is solely determined by the minimum-group-size parameter k, i.e., by ensuring k-anonymity [21], [32].

II. Background

As mentioned in Section I, MAP has been tackled using different techniques. Basically, a MAT relies on a clustering technique and an aggregation technique. MATs were originally used for numerical data [10], [60], and they can further be classified as being *univariate* versus *multivariate* [16]–[18], [24], [25], [32], [34], [44], [47]–[49], *fixed size* versus *data oriented* [13], [24], [25], [45], [47]–[49], [55], [58], and *optimal* versus *heuristic* [18], [39], [53]. Unfortunately, computing the optimal MAP for multivariate microaggregation is an *NP*-hard problem [53]. Therefore, researchers seek heuristic MATs that provide a good solution that is close to the optimal.

The first algorithm that accomplishes microaggregation without projecting the multivariate data onto a single axis was proposed in 2002 by Domingo–Ferrer et.al. [24] and is known as MDAV. It microaggregates the multivariate microdata file based on the concept of the diameter distance of the data set. In 2005, an enhanced version of MDAV appeared in [32] and was implemented as a built-in technique in the μ-ARGUS software tool version 4.0 [41]. The modification is based on utilizing the centroid concept (instead of the diameter) in the microaggregation. The process is briefly explained as follows. First, the algorithm computes the centroid of the data. Then, a quick search for the most distant record from the centroid, e.g., X_r, is done. Subsequently, a new search for the most distant record from the record X_r, e.g., X_s, is accomplished. The next step consists of creating two clusters. The first cluster comprises X_r and *its* $k-1$ nearest records, whereas the second one comprises X_s with *its* nearest $k-1$ records. At the end of this stage, the two clusters are microaggregated and removed from the original data set. The latter steps are iteratively repeated until there no longer are records in the original data set. The advantages of this new modified version of the MDAV are given as follows: 1) an increase in the speed of the microaggregation and 2) reduction in the IL. More recently, the $V-MDAV$ scheme has been proposed to obtain a data-oriented microaggregation solution that provides variable-sized groups and, thus, a higher within-group homogeneity while maintaining an equivalent computational cost [58].

The underlying philosophy that we will use to develop our new scheme relates to the ACNN, which was proposed [66] as a recurrent neural network (NN) model that dynamically evaluates the association of any pair of patterns through the interaction between them and the group of patterns. ACNN possesses several attractive features, e.g., a simple structure, the respective learning mechanism, and an efficient clustering strategy, which uses the association as a new similarity measure. Its superiority in clustering and analyzing gene expression data has also been demonstrated [67]. The rationale behind this superiority probably lies in the inherent advantages of ASRs, which possess the potential to ensure that the similarities between patterns within the same cluster increase, whereas the similarities between different clusters decrease.

ACNN initializes the association between any two neurons by evaluating the relationship between them and by setting the learning ratio α to the most suitable value. The learning ratio should guarantee that the initial association is large when the distance (i.e., proximity in the feature space) between the patterns is small. ACNN studies the interaction level of each pair of patterns based on the association that the other patterns made and defines the similarity threshold, which ensures a robust performance. The association value between any two patterns has to be updated based on the result of the interaction level, and this is, in turn, scaled by using the well-known sigmoid function. This procedure has to iteratively be executed until there is no noticeable change in the successive associations. Subsequently, ACNN constructs the cluster characteristic matrix to describe the cluster property at the end of the learning phase, after which it determines the number of clusters and labels the patterns with their cluster indices.[2]

III. IAMAT

The state-of-the-art MATs use a proximity function, in particular the Euclidean distance, to measure the similarity between records. To the best of our knowledge, the combination of the association and the interaction between individual records has not been taken into consideration while microaggregating the data file. We now discuss how these two criteria are applicable to microaggregating the data file to further minimize the IL.

At the outset, we assert that, although the basic ACNN dynamically evaluates the association and the interaction between the patterns, it is not directly applicable to the MAP in its basic form due to the following four reasons.

1) Unlike the neural setting, in which the weights of the neurons are updated based on the relative relationship between them, it is meaningless to have negative weights for the MAP.
2) To effectively model the switching and clipping effects in complex NN domains, researchers introduced functions, e.g., the sigmoidal function, for transforming the input space by using highly nonlinear mappings. It is our position that such switching and clipping effects are not pertinent to the study of the MAP, because the associations and the interactions between the records are related to their relative proximity, and we have no reason to believe that these quantities *abruptly* fall off or change.
3) The basic ACNN computes the interaction between the neurons by using the two-step product that involves a_{ip} and a_{pj}. Unlike in ACNN, in our solution, we do not compute the sum of all the interactions between the nodes but rather need the one that maximally interacts with the nodes that are already in the same cluster, e.g., X_i and X_j. Thus, as explained in [33], we advocate the use of semiring transitive closure properties similar to a matrix multiplication scheme that is central in determining the multistep Markov matrix for a Markov chain.
4) The final difference between our scheme and ACNN is the fact that we have resorted to a one-shot "training" mechanism,[3] which is *atypical* for most NNs but was used in [1]. Details of this mechanism are also found in [33] and were omitted here for brevity.

Based on the aforementioned principles, we now present the design of our newly proposed IAMAT scheme.

A. Design of IAMAT

We propose that IAMAT aims at microaggregating the records in the data set by using a new methodology for evaluating the similarity between them. This similarity is intuitively expressed by their interrecord relationships and is estimated by measuring their "association" and "interaction" as *modeled* in ACNN. The resulting measurements are similar to the ones that cluster the records based on the distance between them. Consequently, instead of *merely* assigning relatively "close" records to be in the same group, we choose to "estimate" the association and the interaction between them and decide that, if the *combination* of these indices is relatively high, we assign them to be in the same group. Otherwise, we determine that they should be in two different groups. We believe that using this pair of measurements will help us achieve a more robust performance than other existing measures, which is a claim that we have verified and is described as follows.

IAMAT is a consequence of incorporating the aforementioned considerations into the elegant MDAV strategy. Consider IAMAT for any specific value of k. IAMAT uses the centroid of the data set to relatively determine the farthest record, e.g., X_r. Subsequently, we achieve a quick search to obtain the record that is most associated to X_r, e.g., X_s. Then, we propose to choose $k-2$ records based on the mutual *interaction* between each record inside the group and the remaining unassigned records. Consequently, the next step consists of creating a cluster with the associated pair $\langle X_r, X_s \rangle$ and the most interactive

[2]As recommended, the details of ACNN are not included here. We refer the interested reader to [66] for a complete explanation of this clustering strategy. Additional details of these concepts with regard to the MAP are also found in the Ph.D. dissertation of the second author [33], which can be made available to interested readers.

[3]Informally speaking, the associations are computed based on the relative proximities of records, and the interactions are computed based on the latter. Due to this approach, the iterations to recompute the associations only serve to refine the values that were computed in the first iteration. We are interested only in determining whether two patterns are in the same group; thus, we only need a *rough* estimate of the interactions and associations, and thus, further iterative refinement is unnecessary.

$k-2$ records. At the end of this stage, the cluster is microaggregated and removed from the original data set. The aforementioned steps are iteratively repeated until there no longer are $k-1$ records that remain in the original data set. IAMAT terminates by assigning the remaining unassigned records to the last group. The scheme is algorithmically described in Algorithm 1, after which each step is explained in greater detail.

Algorithm 1: IAMAT
Input: The original microdata file $\mathcal{D}$, which contains n unassigned records, and the parameter k.
Output: The microaggregated microdata file $\mathcal{D}'$.
Method:
1: Compute the centroid of $\mathcal{D}$ as $\mu = (1/n)\sum_{i=1}^{n} X_i$.
2: Compute the scaling factor α as related to the mean square distance as $\alpha = \sqrt{n}/(1/n)(\sum_{i=1}^{n} \|X_i - \mu\|^2)$.
3: Compute the association values between μ and each record X_i in $\mathcal{D}$ as $a_{\mu i} = e^{-(\|X_i-\mu\|^2/\alpha)}$.
4: Initialize the number of groups to zero.
5: **while** there are more than $(k-1)$ *unassigned* records in $\mathcal{D}$, **do**
6: Increment the number of groups by unity.
7: Initialize the number of records inside the group to zero.
8: Select the least associated *unassigned* record X_r to the centroid μ as follows: $X_r = Min\ a_{\mu i}$.
9: Mark X_r as an *assigned* record.
10: Compute the association values between X_r and each *unassigned* record X_i in $\mathcal{D}$.
11: Select the most associated *unassigned* record X_s to X_r as follows: $X_s = Max\ a_{ri}$.
12: Mark X_s as an *assigned* record.
13: Compute the association values between X_s and each *unassigned* record X_i in $\mathcal{D}$.
14: Add X_r and X_s to the group and increment the number of records inside the group by two units.
15: **while** the number of records inside the group is less than k, **do**
16: **for all** *unassigned* records X_p in $\mathcal{D}$ **do**
17: Initialize the interaction of X_p, η_p, to 1.
18: **for all** *assigned* records inside the group X_i, **do**
19: Update the value of interaction as follows: $\eta_p = \eta_p * a_{ip}$.
20: **end for**
21: **end for**
22: Let X^* be the record with the highest value for η_p.
23: Mark X^* as an *assigned* record.
24: Add X^* into this group and increment the number of records inside the group by unity.
25: Compute the association values between the most interactive record X^* and each *unassigned* record X_i in $\mathcal{D}$.
26: **end while**
27: Remove the present cluster from the set $\mathcal{D}$.
28: **end while**
29: Assign the remaining *unassigned* records to the last group.
30: Build the microaggregated data file $\mathcal{D}'$.
31: **return** $\mathcal{D}'$.
32: **End Algorithm**: IAMAT

Unlike MDAV, instead of measuring the distance between the records, IAMAT utilizes the association as per ACNN. ACNN classifies the records as being associated if the value of the association index a_{ij} is positive. Otherwise, the neurons will be classified as being unrelated, leading to its "rejection." Clearly, rejecting records will not comply with the spirit and goal of the MAP, whose aim is to minimize the IL. We believe that an association between any pair of records exists, regardless of its value, which can be very small when it is close to zero or is very large when it is close to unity. Therefore, IAMAT quantifies the value of the association between two records, e.g., X_i and X_j, to belong to the interval [0, 1], which is computed as follows:

$$a_{ij} = a_{ji} = r(X_i, X_j) = e^{-\frac{\|X_i - X_j\|^2}{\alpha}} \tag{1}$$

where $r()$ is the identical function that was used in the definition of ACNN, which evaluates the relationship between any two records and also involves α. As mentioned, the value of α is assigned to guarantee that the initial association is large when the distance between X_i and X_j is small, and vice versa. Typically, it is given by

$$\alpha = \frac{\sqrt{n}}{\frac{1}{n}\left(\sum_{i=1}^{n} \left\|X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)\right\|^2\right)}. \tag{2}$$

The rationale for incorporating the association with the interaction between the records inside a group is that it leads to more homogeneous groups. The concept of *interaction* turns out to be crucial in forming the cluster, because we believe that merely being close to the farthest records is not a reason that is sufficiently important for any record to be grouped with the most distant one. Rather, we propose that the interaction with respect to all the records inside the group has to be taken into consideration while clustering the records. As aforementioned, the latter entity is computed by invoking transitive-closure-like operations. Finding the most interactive record with the associated pair is achieved by searching for the maximum product of the association between the unassigned records X_p and each record in the associated pair $\langle X_i, X_j \rangle$ as follows:

$$\eta_{ij} = \begin{cases} a_{ip}(t-1) \times a_{pj}(t-1), & p \neq i, j \quad i \neq j \\ 0, & i = j. \end{cases} \tag{3}$$

The aforementioned equation is valid when $k = 3$. By increasing the value of k, transitive closure is applied by adding one unassigned record at a time. The decision of grouping the unassigned record with other records in the group depends on the interaction of that record with respect to other records inside the group. Logically, the most interactive unassigned record has been chosen as follows:

$$Index\ Maximum_{1 \leq p \leq n}\ \eta_p \tag{4}$$

where

$$\eta_p = \prod_{i=1}^{n_j} a_{ip} \tag{5}$$

where X_i represents the record inside the group G_j of size n_j, and X_p is the unassigned record.

The IAMAT and MDAV methods essentially have the same control structures; thus, once the association and interaction of the records have been computed (which can be done "offline"), both of them will have the same asymptotic time and memory requirements. We now, however, report the experimental results that were obtained by running the two algorithms on simulated and real data sets.

IV. Comparing the MDAV and IAMAT Methods

A. Data Sets

IAMAT has extensively been tested, and the results that were obtained seem to be very good, where the "goodness" of a scheme refers to the combination of its being efficiently computed, and its capability of offering a good trade-off between the IL and the DR. We have tested it using the two real-life benchmark reference data sets that were used in previous studies and in two simulated data sets that were obtained using *Matlab's* built-in functions. These sets are given as follows:

- *Tarragona* data set, which contains 834 records with 13 variables [24];
- *Census* data set, which contains 1080 records with 13 variables [28];
- Uniform distribution $(\min = 0; \max = 40\,000)$, which contains 10 000 16-dimensional records;
- Normal distribution $(\mu = 500; \sigma = 150)$, which contains 6000 records and 16 dimensions.

Note that the simulated data sets have been generated based on a method that is analogous to the method for proving the real-life sets, i.e., by selecting key variables and records [52]. The resulting simulated data sets had the following two properties that are crucial to our experiments.

1) Key variables are necessary to estimate the DR using the Record-Linkage Disclosure (RLD) technique. Therefore, the selection criterion of the key variables was based on choosing the minimum number of repetitions of values in each variable. In particular, three key variables were chosen for the uniform data sets, whereas five key variables were chosen for the normal data set.
2) The number of records in each data set was based on the number of key variables. In general, the size of the simulated data would be limited inasmuch as one would not expect repeated values for continuous variables. However, there were repetitions in the data set. After the repetitions were deleted, our selection of 4800 records in the uniform data set and 1560 in the normal data set resulted, because this value was the cardinality of the set that corresponds to the largest integer, which is a multiple of 3, 4, 5, and 6. Thus, for comparison, the MAT could be invoked with a minimum group size of either $k = 3, 4, 5,$ or 6.

To further investigate the performance of the new scheme, several experiments were conducted using various simulated data sets, that involve *independent* vectors with dimensions that range from 10 to 80 and sets of cardinality that range from 10 000 to 100 000, which were generated using *Matlab's* built-in-functions for the following two types of distribution: 1) uniform $(\min = 0; \max = 1000)$ and 2) normal $(\mu = 0; \sigma = 0.05)$.

B. Comparing the MDAV and IAMAT Methods

Quantifying the quality of MATs is based on two criteria, i.e., the IL and the DR, both of which have been explained in [28] and [65]. In this paper, we have also considered how we can compare MATs using a *composite* measure that involves *both* the IL and the DR.

1) IL: As mentioned in [9], [26], [51], [52], and [57], several measures for quantifying the IL have been proposed. Five of these measures[4] will be used in this paper to construct a comparative benchmark. This will be done by assuming that the original and masked microdata sets are specified in terms of the n-ordered individuals, e.g., $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ and $\mathcal{X}' = \{X'_1, X'_2, \ldots, X'_n\}$, respectively. Observe that each X_i is an instantiation of the random vector (of dimension d) $\underline{X}$, whose mean is $\bar{X}$, and each X'_i is an instantiation of the random variable $\underline{X'}$, whose mean is $\bar{X}'$. Thus, each data vector in the original and masked data sets can be represented as $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T$ and $X'_i = [x'_{i1}, x'_{i2}, \ldots, x'_{id}]^T$, respectively, where both x_{ij} and x'_{ij} are the values that are associated with the jth variable. Thus, we symbolically use the following notation.

$\mathcal{X}$ and $\mathcal{X}'$	Original and masked data sets, respectively.
$\bar{X}$ and $\bar{X}'$	Mean vectors of $\underline{X}$ and $\underline{X'}$, respectively, which are computed as $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ and $\bar{X}' = (1/n)\sum_{i=1}^{n} X'_i$.
$\mathcal{V}$ and $\mathcal{V}'$	Covariance matrices of $\underline{X}$ and $\underline{X'}$, respectively; thus, $\mathcal{V} = \mathrm{E}[(\underline{X} - \bar{X})(\underline{X} - \bar{X})^T]$, and $\mathcal{V}' = \mathrm{E}[(\underline{X'} - \bar{X}')(\underline{X'} - \bar{X}')^T]$.
S and S'	Vectors that represent the variance of the components of $\underline{X}$ and $\underline{X'}$, respectively, and are given as $S = \mathrm{Diag}[\mathcal{V}]$ and $S' = \mathrm{Diag}[\mathcal{V}']$.
$\mathcal{R}$ and $\mathcal{R}'$	Correlation matrices of $\underline{X}$ and $\underline{X'}$, respectively; if Γ is the diagonal matrix with the standard deviation of the variables along the main diagonal (and with zeros elsewhere), then $\mathcal{R} = \Gamma^{-1}\mathcal{V}\Gamma^{-1}$, and $\mathcal{R}' = \Gamma'^{-1}\mathcal{V}'\Gamma'^{-1}$.

To quantify the difference or the "discrepancy" between two matrices, in this paper, we use the mean variation[5] for data structures, means, variances, and covariances, whereas the mean absolute error is used to measure the data correlation difference, as shown in Table I. This leads to five distinct metrics, i.e., $M1$–$M5$, whose significance and explicit forms

[4]The rationale for these measures and the measures for quantifying the DR was better explained in [33], which can be made available to interested readers.

[5]The following rule is applied to all the mean variation formulas: if $x_{ij} = 0$ and $x'_{ij} \neq 0$, then we divide the difference by $|x'_{ij}|$, and if $x_{ij} = x'_{ij} = 0$, the term is not added to the sum.

TABLE I
IL MEASURES

Metric	Measurement	Matrix discrepancy	Mathematical expression
$M1$	Mean variation of data	$\mathcal{X} - \mathcal{X}'$	$\frac{\sum_{j=1}^{d} \sum_{i=1}^{n} \frac{\lvert x_{ij} - x'_{ij} \rvert}{\lvert x_{ij} \rvert}}{nd}$
$M2$	Mean variation of data means	$\bar{X} - \bar{X}'$	$\frac{\sum_{j=1}^{d} \frac{\lvert \bar{x}_j - \bar{x}'_j \rvert}{\lvert \bar{x}_j \rvert}}{d}$
$M3$	Mean variation of data variances	$S - S'$	$\frac{\sum_{j=1}^{d} \frac{\lvert v_{jj} - v'_{jj} \rvert}{\lvert v_{jj} \rvert}}{d}$
$M4$	Mean variation of data covariates	$\mathcal{V} - \mathcal{V}'$	$\frac{\sum_{j=1}^{d} \sum_{1 \le i \le j} \frac{\lvert v_{ij} - v'_{ij} \rvert}{\lvert v_{ij} \rvert}}{\frac{d(d+1)}{2}}$
$M5$	Mean absolute error of data correlations	$\mathcal{R} - \mathcal{R}'$	$\frac{\sum_{j=1}^{d} \sum_{1 \le i < j} \lvert r_{ij} - r'_{ij} \rvert}{\frac{d(d-1)}{2}}$

are tabulated in Table I. Finally, the overall IL, G_{IL} is defined as follows:

$$G_{\mathrm{IL}} = 100 * \frac{M1 + M2 + M3 + M4 + M5}{5}. \tag{6}$$

2) DR: The effect of applying different MATs is not, in practice, limited to the IL. Rather, we need to evaluate the risk that the values of original records can accurately be estimated from the published masked records [7], [51], [63]. The first kind of risk is evaluated through the RLD technique [11], [23], [26], [28], and the second kind is evaluated through the confidential Interval Disclosure (ID) technique[6] [11], [23], [26]. The overall global DR, G_{DR} is defined as the average of the DR as computed by both aforementioned strategies and has the form

$$G_{\mathrm{DR}} = \frac{\mathrm{RLD} + \mathrm{ID}}{2}. \tag{7}$$

3) Overall SI: As we know, the indices of the IL and the DR reflect completely different aspects of a MAT. As argued earlier, we feel that a fairer index would be one that simultaneously considers both of them. In this context, we define the SI to be a linear combination (with $x = 0.5$) of the indices, which was obtained for the IL and the DR as

$$\mathrm{SI} = x\, G_{\mathrm{IL}} + (1 - x)\, G_{\mathrm{DR}}. \tag{8}$$

C. Results

For a given value of the security parameter k, which represents the minimum number of records per group, we compared the percentage value of IL = (SSE/SST) (as defined in Section I) that results from the IAMAT and MDAV strategies. Note that MDAV was implemented based on the centroid concept and not on a diameter concept.[7] All the programs were written in the C++ language, and the tests were performed on a 1.73-GHz Intel(R) Pentium (R)M processor with 512-MB *RAM*.

Table II shows the improvement of the solution that was obtained using IAMAT compared to MDAV on the multivariate real data sets, where all the 13 variables were simultaneously used during the microaggregation process. We attained a reduction in the value of the IL of up to 8% on the Tarragona data set and 5.12% on the Census data set when the group size was equal to 3. However, in the case of the simulated data sets, the improvement in IL reached up to 14.47% when $k = 5$ in the uniform data set, and it was as high as 12.58% when $k = 5$ in the normal data set. Thus, it is evident that the impact of the group size on the solution is minimized by increasing the number of records per group. To be fair, the computational time for executing IAMAT is almost double the computational time for MDAV, although in every case, the time was marginal, i.e., less than 0.5 s. In terms of comparison, we believe that minimizing the loss in the data utility is more important than minimizing the extremely small computational time, particularly because microaggregation is usually performed offline. However, the question of how the decrease in IL is related to the increase in the computational time is still open. In addition, we are not able to explain why the improvement in Table II somewhat increases with k for some data sets and decreases for others, which is probably due to the peculiarity of the specific data sets.

TABLE II
COMPARISON OF THE PERCENTAGE OF THE IL AND THE COMPUTATIONAL TIME BETWEEN MDAV AND IAMAT ON THE REAL-LIFE DATA SETS (TARRAGONA AND CENSUS) AND THE SIMULATED DATA SETS (UNIFORM AND NORMAL DISTRIBUTIONS) FOR MULTIVARIATE METHODS

Data Set	k value	*MDAV* IL	*MDAV* Time	*IAMAT* IL	*IAMAT* Time	Improv. (%)
Tarragona	3	16.9593	0.17	15.6023	0.31	8.00
	4	19.7482	0.12	19.2872	0.22	2.33
	5	22.8850	0.12	22.7164	0.23	0.74
Census	3	5.6535	0.22	5.3639	0.41	5.12
	4	7.4414	0.19	7.2170	0.44	3.02
	5	8.8840	0.17	8.8428	0.42	0.46
	6	10.1941	0.17	9.9871	0.42	2.03
Uniform Distribution	3	22.4608	4.18	19.9730	8.52	11.08
	4	29.1714	4.45	25.2008	7.85	13.61
	5	33.9636	4.62	29.0478	7.48	14.47
	6	37.1577	4.75	32.1441	7.18	13.49
Normal Distribution	3	26.9348	0.45	24.1245	0.92	10.43
	4	33.9256	0.78	30.0154	0.81	11.53
	5	39.564	0.59	34.5886	0.87	12.58
	6	43.4592	0.60	38.0086	0.75	12.54

The other experiments were carried to test the SI of MDAV and IAMAT. Therefore, they have been scored with respect to

[6]The details and relevance of these risks are not included here. They are well described in the literature, and additional details of their computation and relevance to the MAP are also found in [33].

[7]We did not program the MDAV scheme. We are extremely thankful to Dr. F. Sebé for giving us his source code.

TABLE III
SCORING MDAV AND IAMAT WITH RESPECT TO $G_{\rm IL}$ AND $G_{\rm DR}$ BY COMPUTING THE INDEX SI, FOR $k = 3, 4$, AND 5, BY USING THE CENSUS DATA SET AND THE SIMULATED UNIFORMLY AND NORMALLY DISTRIBUTED DATA SETS

Data	*Creterion*	$k=3$		$k=4$		$k=5$	
		MDAV	*IAMAT*	*MDAV*	*IAMAT*	*MDAV*	*IAMAT*
Census	G_{IL}	28.6100	22.1500	33.2540	30.3120	38.6720	34.9580
	RLD	60.7133	61.8889	49.7057	51.9070	42.0981	43.1771
	ID	1.9815	1.8704	0.7315	0.8611	0.2685	0.5556
	G_{DR}	31.3474	31.8796	25.2186	26.3840	21.1833	21.8663
	SI	33.8050	30.5750	36.1270	34.6560	38.8360	36.9790
Uniform	G_{IL}	115.4319	37.8225	116.1198	45.1763	120.9199	53.3052
	RLD	0.1088	1.1875	0.0694	0.8449	0.0602	0.5556
	ID	0.0042	4.7194	0.0000	1.2437	0.0000	0.4924
	G_{DR}	0.0565	2.9535	0.0347	1.0443	0.0301	0.5240
	SI	57.7442	20.3880	58.0773	23.1103	60.4750	26.9146
Normal	G_{IL}	108.9812	61.8348	89.0398	80.1520	110.8694	105.2394
	RLD	0.6494	11.8173	0.6571	6.7410	0.2814	5.3885
	ID	0.0085	0.8248	0.0000	0.1987	0.0000	0.0406
	G_{DR}	0.3290	6.3210	0.3285	3.4699	0.1407	2.7145
	SI	54.6551	34.0779	44.6841	41.8109	55.5051	53.9770

the SI index on the simulated data and the Census[8] data sets, which contains the following seven key variables:

1) *FEDTAX*;
2) *AFNLWGT*;
3) *AGI*;
4) *EMCONTRB*;
5) *PTOTVAL*;
6) *TAXING*;
7) *STATETTAX*.

Table III displays the SI for the MDAV and IAMAT methods for various values of k, which was set to be either 3, 4, or 5 based on the accepted requirements in [25], [55], and [62]. Based on (6), the $G_{\rm IL}$ was computed by averaging the values of M_1–M_5. In general, the value of $G_{\rm IL}$ is "directly" proportional to the number of records per group k. Therefore, in the Census data set, the best value of $G_{\rm IL}$ for IAMAT was obtained when $k = 3$ and was equal to 22.15%, whereas the best value of $G_{\rm IL}$ for MDAV was 28.61% when $k = 3$. In terms of the simulated data sets (and in general), the value of $G_{\rm IL}$ for IAMAT is less than half the value that was obtained using MDAV (i.e., the value of $G_{\rm IL}$ for IAMAT in the uniform data set and when $k = 3$ was 37.82%, whereas it was equal to 115.43% for MDAV). Clearly, the IAMAT method more efficiently preserves data utility than the state of the art, i.e., MDAV. The table also shows a comparison of $G_{\rm DR}$ (which estimates the risk of the data being disclosed) using the RLD and ID techniques. The superiority of IAMAT is clear. Furthermore, generally speaking, the results show that estimating the risk of disclosing the secure information using the RLD method falls inversely "proportional" to the number of records per group k.

Finally, the SI value was computed for each MAT based on (8). Observe that a lower score value implies a superior performance. In the table, we see that the IAMAT technique has, almost consistently, better performance index than MDAV based on not only the $G_{\rm IL}$ perspective but also the perspective of a combination of the $G_{\rm IL}$ and $G_{\rm DR}$ for different values of k.

[8]Scoring them against the Tarragona data set is meaningless, because the latter data set does not contain the so-called key variables.

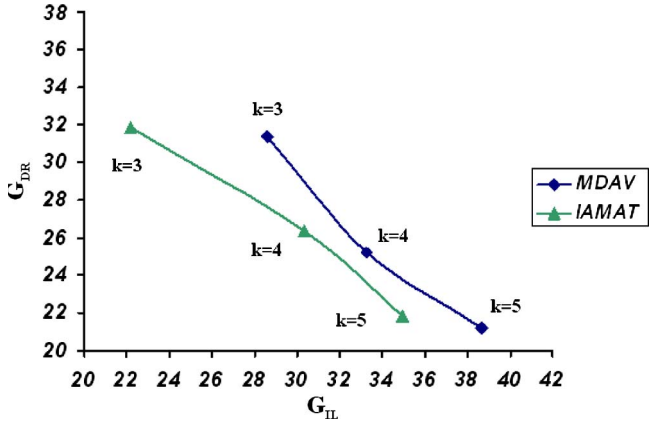

Fig. 1. Effect of invoking MDAV and IAMAT on the $G_{\rm IL}$ and $G_{\rm DR}$ indices when $k = 3, 4$, and 5 for the Census data set.

Thus, for example, in the Census data set, the IAMAT method scores a minimum value of 30.58 when $k = 3$, 34.66 when $k = 4$, and 36.98% when $k = 5$. In addition, IAMAT scores almost half the SI value, which was obtained by MDAV by using the simulated uniform and normal data sets.

To investigate how IAMAT and MDAV compare with respect to "all" the factors, in Fig. 1, we have plotted $G_{\rm IL}$ versus $G_{\rm DR}$ for the Census data set for both schemes. This figure presents sets of paired values of $G_{\rm IL}$ and $G_{\rm DR}$ for the respective algorithm for different values of k, which range from 3 to 5. Based on the curve, IAMAT is shown to optimize these conflicting criteria in a superior way than the state-of-the-art MADV method, because as the value of k increases, the increase in $G_{\rm IL}$ does not affect the value of $G_{\rm DR}$ in the IAMAT compared to the case of MDAV.

We also undertook a comprehensive evaluation of the performance of the IAMAT scheme to investigate the *scalability* of the technique with respect to the cardinality of the data set, its dimensionality, and the number of records per group. Although the details of these results are omitted here for brevity,[9] more detailed results of every one of these scenarios are found in

[9]These results were presented in a concise manner as per recommendation.

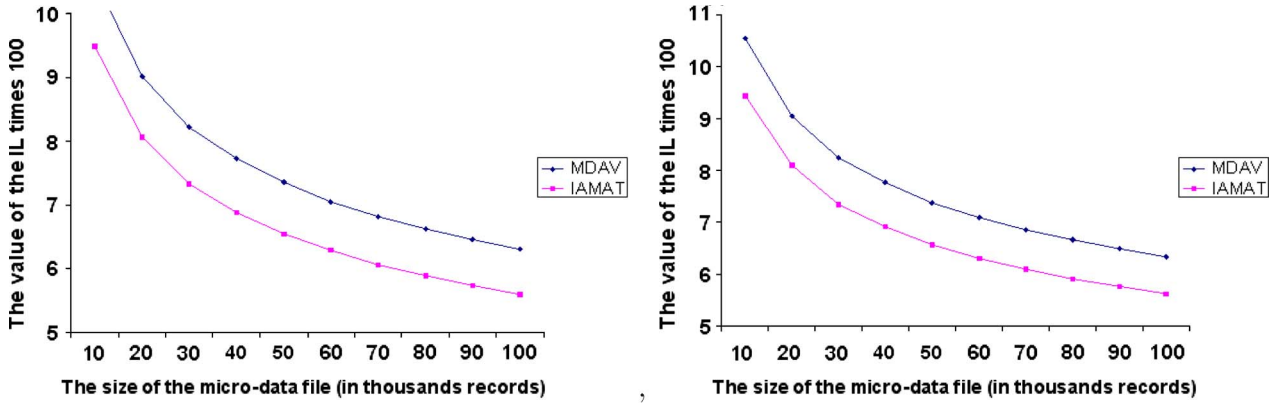

Fig. 2. Improvement of IAMAT in reducing the percentage value of the IL as a function of the cardinality of the data set for (left) normally and (right) uniformly distributed data. In both cases, $k = 3$.

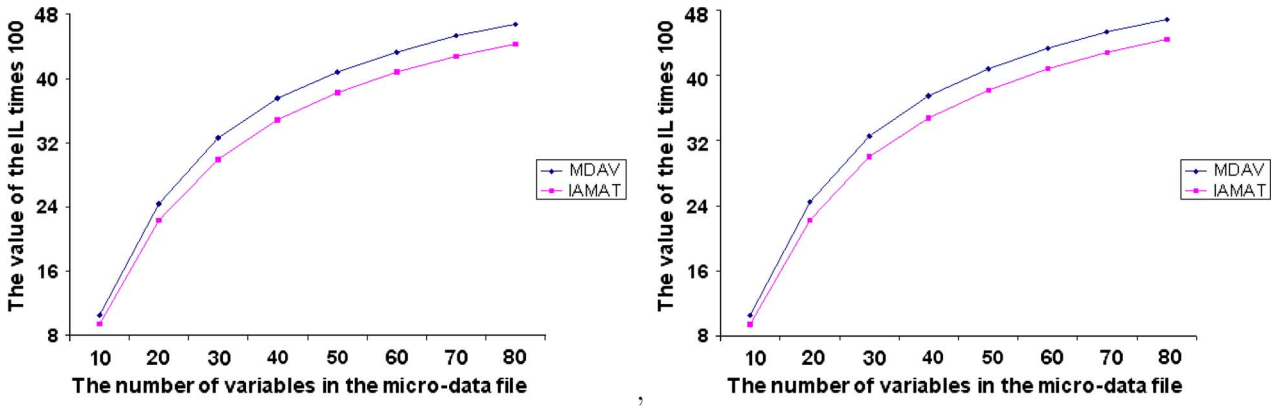

Fig. 3. Improvement of IAMAT in reducing the percentage value of the IL as a function of the dimension of the data set for (left) normally and (right) uniformly distributed data. In both cases, $k = 3$, and $n = 10\,000$.

[33]. In summary, with respect to the **scalability of IAMAT with respect to cardinality**, it is fair to state that IAMAT was superior to MDAV in every single case. For example, the percentage of improvement that was achieved by invoking IAMAT in the IL (when the value of k was set to 3) ranged from 10.02% to 11.25% for the normal distribution and from 10.47% to 11.28% for the uniform distribution. Furthermore, in general, increasing the size of the data set tends to minimize the IL value, with the IAMAT being always the superior scheme, as shown in Fig. 2. With respect to the **scalability of IAMAT with respect to dimensionality**, as expected, increasing the dimensionality implies increasing the loss in data utility. Again, IAMAT was superior, and the highest percentage of the improvement in the IL was about 10% for both the uniform and normal distributions. Fig. 3 shows that both the IAMAT and MDAV schemes have similar trends for the IL i.e., they are almost proportional to the dimension of the variables in the microaggregation process for both distributions. Finally, with respect to the **scalability of IAMAT with respect to the number of records per group**, IAMAT was always the superior method, where the reduction in the IL reached 12.74% for the normal distribution when the group size was 5 and 12.31% for the uniform distribution when the group size was 4. Fig. 4 shows that both schemes possess similar trends for the IL as a function of the group size and that the value of the IL is proportional to the number of records per group.

V. Conclusion

In this paper, we have considered the problem of achieving microaggregation in secure statistical databases. The novelty of our method involves enhancing the primitive MAT, which merely incorporates proximity information. The state-of-the-art MAT recursively reduces the size of the data set by excluding points that were farthest from the centroid and points that were closest to these farthest points. Thus, although the state-of-the-art method was extremely effective, we have argued that it uses only the proximity information and ignores the mutual interaction between the records. In this paper, we have proven that interrecord relationships can be quantified in terms of two entities, i.e., their "association" and "interaction," which can be measured by invoking transitive-closure-like operations, and by mapping the problem into a neural setting using ACNN. By repeatedly invoking the interrecord associations and interactions, we have shown that the records can be grouped into sizes of cardinality "k." Our experimental results, which were done on artificial data and on the benchmark data sets for real-life data, demonstrate that the newly proposed method is superior to the state of the art by as much as 13%.

By defining a score SI as a composite measure that involves the IL and the DR, it has been shown that the proposed strategy also obtains a minimum score value compared to the MDAV method. This result indicates that the IAMAT technique is, probably, the best MAT based on not only the IL perspective

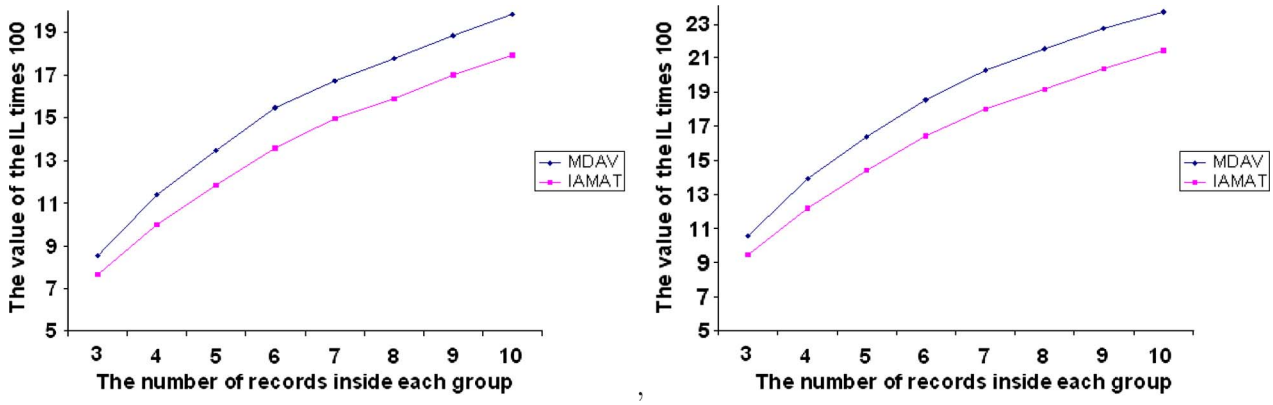

Fig. 4. Improvement of IAMAT in reducing the percentage value of the IL as a function of the number of records per group for (left) normally and (right) uniformly distributed data. In both cases, $n = 10\,000$, and $d = 10$.

but also the viewpoint of a measure, i.e., as a combination of the IL and DR.

We foresee two avenues for future work: 1) extending IAMAT toward data-oriented microaggregation, where the group size n_i satisfies $k \leq n_i < 2k$, and 2) investigating the effect of having a dynamic value of α on the compactness of each group and on the value of the IL.

ACKNOWLEDGMENT

The authors would like to thank Dr. Josep Domingo–Ferrer for all his support and advice, for providing the data sets, and, along with Dr. Josep Mateo–Sanz and Dr. Francesc Sebé, for answering our queries and providing the code for the MDAV scheme, and the anonymous Referees for their valuable comments, which significantly improved the quality of this paper.

REFERENCES

[1] M. Adachi and K. Aihara, "Associative dynamics in a chaotic neural network," *Neural Netw.*, vol. 10, no. 1, pp. 83–98, Jan. 1997.
[2] N. Adam and J. Wortmann, "Security-control methods for statistical databases: A comparative study," *ACM Comput. Surv.*, vol. 21, no. 4, pp. 515–556, Dec. 1989.
[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD*, Washington, DC, 1993, pp. 207–216.
[4] R. Agrawal, H. Mannila, H. Srikant, R. Toivonen, and I. Verkamo, "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, 1996, pp. 307–328.
[5] F. Bacao, V. Lobo, and M. Painho, "Self-organizing maps as substitutes for K-means clustering," in *Proc. Int. Conf. Comput. Sci.*, 2005, pp. 476–483.
[6] Y. Baeyens and D. Defays, "Estimation of variance loss following microaggregation by the individual ranking method," in *Proc. Stat. Data Protection*, 1998, pp. 101–108.
[7] R. Brand, J. Domingo-Ferrer, and J. Mateo-Sanz. (2002). "Reference data sets to test and compare SDC methods for protection of numerical microdata," Tech. Rep., CASC PROJECT, Computational Aspects of Statistical Confidentiality. [Online]. Available: http://neon.vb.cbs.nl/casc/CASCrefmicrodata.pdf
[8] G. Crises, "An introduction to microdata protection for database privacy," Tech. Rep. CRIREP-04-006, 2004.
[9] G. Crises, "Information loss measures for microdata in database privacy protection," Tech. Rep. CRIREP-04-004, 2004.
[10] G. Crises, "Microaggregation for privacy protection in statistical databases," Tech. Rep. CRIREP-04-005, 2004.
[11] G. Crises, "Microdata disclosure risk in database privacy protection," Tech. Rep. CRIREP-04-003, 2004.
[12] G. Crises, "Trading off information loss and disclosure risk in database privacy protection," Tech. Rep. CRIREP-04-002, 2004.
[13] M. Cuppen, "Secure data perturbation in statistical disclosure control," Ph.D. dissertation, Statistics Netherlands, Voorburg, The Netherlands, 2000.
[14] R. Dandekar, M. Cohen, and N. Kirkendall, "Sensitive microdata protection using the Latin hypercube sampling technique," in *Inference Control in Statistical Databases: From Theory to Practice*. London, U.K.: Springer-Verlag, 2002, pp. 117–125.
[15] C. Date, *An Introduction to Database Systems*. Reading, MA: Addison-Wesley, 2000.
[16] D. Defays and M. Anwar, "Masking microdata using microaggregation," *J. Official Statist.*, vol. 14, no. 4, pp. 449–461, Dec. 1998.
[17] D. Defays and N. Anwar, "Microaggregation: A generic method," in *Proc. 2nd Int. Symp. Stat. Confidentiality*, 1995, pp. 69–78.
[18] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: The small aggregates method," in *Proc. Symp. Des. Anal. Longitudinal Surv.*, 1992, pp. 195–204.
[19] D. Denning, "Secure statistical databases with random sample queries," *ACM Trans. Database Syst.*, vol. 5, no. 3, pp. 291–315, Sep. 1980.
[20] J. Domingo-Ferrer. (1999). "Statistical disclosure control in Catalonia and the CRISES group," Tech. Rep. [Online]. Available: http://vneumann.etse.urv.es/publications/nonsci/questio26-3.pdf
[21] J. Domingo-Ferrer, "Microaggregation: Achieving k-anonymity with quasioptimal data quality," in *Proc. Q: Eur. Conf. Quality Surv. Statist.*, 2006. [Online]. Available: www.statistics.gov.uk/events/q2006/downloads/T06_Ferrer.doc
[22] J. Domingo-Ferrer, A. Martínez-Ballesté, J. Mateo-Sanz, and F. Sebé, "Efficient multivariate data-oriented microaggregation," *Int. J. Very Large Databases*, vol. 15, no. 4, pp. 355–369, Sep. 2006.
[23] J. Domingo-Ferrer and J. Mateo-Sanz, "An empirical comparison of SDC methods for continuous microdata in terms of information loss and disclosure risk," in *Proc. Joint ECE/Eurostat Work Session Stat. Data Confidentiality, Conf. Eur. Satisticians*, 2001.
[24] J. Domingo-Ferrer and J. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, Jan./Feb. 2002.
[25] J. Domingo-Ferrer, J. Mateo-Sanz, A. Oganian, V. Torra, and A. Torres, "On the security of microaggregation with individual ranking: Analytical attacks," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 477–491, Oct. 2002.
[26] J. Domingo-Ferrer, J. Mateo-Sanz, and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk," in *Proc. ETK-NTTS*, 2001, vol. 2, pp. 807–825.
[27] J. Domingo-Ferrer and F. Sebé, "Optimal multivariate 2-microaggregation for microdata protection: A 2-approximation," in *Proc. Privacy Stat. Databases*, Rome, Italy, 2006, pp. 129–138.
[28] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, Eds. Amsterdam, The Netherlands: North Holland, 2002, pp. 113–134.
[29] J. Domingo-Ferrer and V. Torra, "Aggregation techniques for statistical confidentiality," in *Aggregation Operators: New Trends and Applications*. Heidelberg, Germany: Physica-Verlag GmbH, 2002, pp. 260–271.
[30] J. Domingo-Ferrer and V. Torra, "Disclosure control methods and information loss for microdata," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, The Netherlands: North Holland, 2002, pp. 93–112.

[31] J. Domingo-Ferrer and V. Torra, "Fuzzy microaggregation for microdata protection," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 7, no. 2, pp. 153–159, Jun. 2003.
[32] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous, and heterogeneous k-anonymity through microaggregation," *Data Mining Knowl. Discovery*, vol. 11, no. 2, pp. 195–212, Sep. 2005.
[33] E. Fayyoumi, "Novel microaggregation techniques for secure statistical databases," Ph.D. dissertation, School Comput. Sci., Carleton Univ., Ottawa, ON, Canada, 2008.
[34] E. Fayyoumi and B. J. Oommen, "Achieving microaggregation for secure statistical databases using fixed structure partitioning-based learning automata," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 5, pp. 1192–1205, Oct. 2009.
[35] F. Felso, J. Theeuwes, and G. Wagner, "Disclosure limitation methods in use: Results of a survey," in *Confidentiality, Disclosure, and Data Access*. Amsterdam, The Netherlands: North Holland, 2001, pp. 17–42.
[36] L. Feng, T. Dillon, H. Weigana, and E. Chang, "An XML-enabled association rule framework," in *Proc. DEXA*, Czech Republic, Prague, 2003, pp. 88–97.
[37] E. Fernandez, R. Summers, and C. Wood, *Databases Security and Integrity*. Reading, MA: Addison-Wesley, 1980.
[38] S. Giessing and A. Hundepool. (2003). "The CASC project: Integrating best practice methods for statistical confidentiality," Tech. Rep. [Online]. Available: http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/79.pdf
[39] S. Hansen and S. Mukherjee, "A polynomial algorithm for univariate optimal microaggregation," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 1043–1044, Jul./Aug. 2003.
[40] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Nordholt, G. Seri, and P. Wolf, *Handbook on Statistical Disclosure Control*. CENEX SDC, 2006.
[41] A. Hundepool, A. Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P. Wolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing, *M-ARGUS Version 4.0 Software and User's Manual*. The Hague, The Netherlands: Statistics Netherlands, 2004.
[42] W. Jonge, "Compromising statistical databases responding to queries about means," *ACM Trans. Database Syst.*, vol. 8, no. 1, pp. 60–80, Mar. 1983.
[43] J. Kim and W. Winkler, "Masking microdata files," in *Proc. Section Surv. Res. Methods*, 1995, pp. 114–119.
[44] M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 7, pp. 902–911, Jul. 2005.
[45] Y. Li, S. Zhu, L. Wang, and S. Jajodia, "A privacy-enhanced microaggregation method," in *Proc. 2nd Int. Symp. FoIKS*, 2002, pp. 148–159.
[46] M. Markey, J. Lo, G. Tourassi, and C. Floyd, Jr., "Self-organizing map for cluster analysis of a breast cancer database," *Artif. Intell. Med.*, vol. 27, no. 2, pp. 113–127, Feb. 2003.
[47] M. Mas, "Statistical data protection techniques," Eustat: Euskal Estatistika Erakundea, Instituto Vasco De Estadistica, Vitoria-Gasteiz, Spain, 2006. Tech. Rep.
[48] J. Mateo-Sanz and J. Domingo-Ferrer, "A comparative study of microaggregation methods," *Questiio*, vol. 22, no. 3, pp. 511–526, 1998.
[49] J. Mateo-Sanz and J. Domingo-Ferrer, "A method for data-oriented multivariate microaggregation," in *Proc. Stat. Data Protection*, 1998, pp. 89–99.
[50] J. Mateo-Sanz, J. Domingo-Ferrer, and F. Sebé, "Probabilistic information loss measures in confidentiality protection of continuous microdata," *Data Mining Knowl. Discovery*, vol. 11, no. 2, pp. 181–193, Sep. 2005.
[51] J. Mateo-Sanz, F. Sebé, and J. Domingo-Ferrer, "Outlier protection in continuous microdata masking," in *Proc. Privacy Stat. Databases*, Barcelona, Spain, 2004, pp. 201–215.
[52] A. Oganian, "Security and information loss in statistical data protection," Ph.D. dissertation, Univ. URV Catalunya, Tarragona, Spain, 2002.
[53] A. Oganian and J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control," *Stat. J. United Nations Econ. Commission Eur.*, vol. 18, no. 4, pp. 345–354, 2001.
[54] B. J. Oommen and E. Fayyoumi, "A novel method for microaggregation in secure statistical databases using association and interaction," in *Proc. 9th Int. Conf. Inf. Commun. Security*. New York: Springer-Verlag, 2007, vol. 4861, pp. 126–140.
[55] J. Panaretos and N. Tzyvidis, "Aspects of estimation procedures at Eurostat with some emphasis on overspace harmonization," in *Proc. HERCMA Conf.*, 2001, pp. 853–857.
[56] J. Park, M.-S. Chen, and P. Yu, "Using a hash-based method with transaction trimming for mining association rules," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 5, pp. 813–826, Sep./Oct. 1997.
[57] J. Sanchez, J. Urrutia, and E. Ripoll. (2003). "Test report on multivariate microaggregation in m-Argus 3.2," Tech. Rep., CASC PROJECT, Computational Aspects of Statistical Confidentiality. [Online]. Available: http://neon.vb.cbs.nl/casc/deliv/6-D6TestreportQuantMA.pdf
[58] A. Solanas and A. Martínez-Ballesté, "V-MDAV: A multivariate microaggregation with variable group size," in *Proc. 17th COMPSTAT Symp. IASC*, Rome, Italy, 2006.
[59] A. Solanas, A. Martínez-Ballesté, J. Mateo-Sanz, and J. Domingo-Ferrer, "Multivariate microaggregation-based genetic algorithms," in *Proc. 3rd Int. IEEE Conf. Intell. Syst.*, 2006, pp. 65–70.
[60] V. Torra, "Microaggregation for categorical variables: A median-based approach," in *Proc. PSD: CASC Project Int. Workshop*, J. Domingo-Ferrer and V. Torra, Eds., Barcelona, Spain, 2004, pp. 162–174.
[61] V. Torra and J. Domingo-Ferrer, "Towards fuzzy C-means-based microaggregation," in *Advances in Soft Computing: Soft Methods in Probability, Statistics, and Data Analysis*, P. Grzegorzewski, O. Hryniewicz, and M. Gil, Eds. Heidelberg, Germany: Physica-Verlag, 2002, pp. 289–294.
[62] T. Wende, "Different grades of statistical disclosure control correlated with German Statistics Law," in *Proc. Privacy Stat. Databases*, Barcelona, Spain, 2004, pp. 336–342.
[63] L. Willenborg and T. Waal, *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001. ILL Number: 2132712.
[64] W. Winkler, "Reidentification methods for masked microdata," in *Proc. PSD: CASC Project Int. Workshop*, J. Domingo-Ferrer and V. Torra, Eds., Barcelona, Spain, 2004, pp. 216–223.
[65] W. Yancey, W. Winkler, and R. Creecy, "Disclosure risk assessment in perturbative microdata protection," in *Inference Control in Statistical Databases: From Theory to Practice*. London, U.K.: Springer-Verlag, 2002, pp. 135–152.
[66] Y. Yao, L. Chen, and Y. Chen, "Associative clustering for clusters of arbitrary distribution shapes," *Neural Process. Lett.*, vol. 14, no. 3, pp. 169–177, Dec. 2001.
[67] Y. Yao, L. Chen, A. Goh, and A. Wong, "Clustering gene data via associative clustering neural network," in *Proc. 9th ICONIP*, 2002, vol. 5, pp. 2228–2232.

B. John Oommen (F'03) was born in Coonoor, India, on September 9, 1953. He received the B.Tech. degree from the Indian Institute of Technology, Madras, India, in 1975, the M.E. degree from the Indian Institute of Science, Bangalore, India, in 1977, and the M.S. and Ph.D. degrees from Purdue University, West Lafayettte, IN, in 1979 and 1982, respectively.

In the academic year 1981–1982, he joined the School of Computer Science, Carleton University, Ottawa, ON, Canada, where he is currently a Full Professor. He is the author of more than 310 refereed journal publications and conference proceedings. He has been on the Editorial Board of *Pattern Recognition*. His research interests include automata learning, adaptive data structures, statistical and syntactic pattern recognition, stochastic algorithms, and partitioning algorithms.

Dr. Oommen is a Fellow of the International Association for Pattern Recognition. He has been on the Editorial Board of the IEEE TRANSACTIONS ON SYSTEMS, MAN, and CYBERNETICS. He received the honorary rank of Chancellor's Professor, which is a lifetime award, from Carleton University in July 2006.

Ebaa Fayyoumi was born in Kuwait in 1978. She received the B.Sc. degree from the Hashemite University, Zarqa, Jordan, in 2000, the M.SC. degree from the University of Jordan, Amman, Jordan, in 2002, and the Ph.D. degree from Carleton University, Ottawa, ON, Canada, in 2008.

Since November 2008, she has been with the Faculty of Prince Hussein Bin Abdalla II for Information Technology, Hashemite University, where she is currently an Assistant Professor and the Chair of the Department of Computer Information System. Her research interests include microaggregation techniques, secure statistical databases, and data mining.